BREAST CANCER DETECTION USING MACHINE LEARNING TECHNIQUES

Harnitha Suresh[1], Shruti Sundaram[2], Kunal Choudhary[3], Tejas Chavan[4], Omkar Mainkar[5]

[1-5]*Student, Department of Computer Science, University of Nottingham, Nottingham, England*

ABSTRACT

The overarching objective of this study revolves around employing advanced machine learning methodologies to predict pathological complete response (PCR) and relapse-free survival (RFS) in breast cancer patients engaging in chemotherapy, leveraging a dataset which comprises clinically measured features as well as pre-chemotherapy MRI-derived features. Stringent validation procedures will ensure generalization to unexplored data, addressing performance, interpretability, and ethical considerations. The research aims to enhance breast cancer treatment possibilities, focusing on personalized and effective approaches for diverse patient demographics.

1. INTRODUCTION

Breast cancer remains a prominent health concern for women in the United Kingdom, necessitating effective treatment strategies. Chemotherapy is commonly employed to shrink locally advanced malignancies before surgical intervention, yet its variable effectiveness and intrinsic toxicity pose challenges. Achieving pathological complete response (PCR) post-surgery enhances the probability of cure and prolongs relapse-free survival (RFS). Despite chemotherapy's critical role, only 25% of patients achieve PCR, leaving 75% with residual disease and diverse prognoses. Improved patient segmentation and tailored treatment are imperative. Advanced machine learning algorithms, utilizing clinically assessed parameters and features extracted from pre-chemotherapy magnetic resonance imaging (MRI), offer a promising avenue for predictive competence in estimating PCR and RFS.

2. DATASET

The dataset for this research is sourced from The American College of Radiology Imaging Network's I-SPY 2 TRIAL, dedicated to breast cancer research. The training dataset (trainDataset.xls) comprises 400 patients with 10 clinical features, including Age, ER, PgG, HER2, Triple-Negative Status, Chemotherapy Grade, Tumor Proliferation, Histology Type, Lymph Node Status, and Tumor Stage. Additionally, 107 MRI-based characteristics obtained using Pyradiomics elevate the dataset's complexity. The value "999" denotes missing data points, reflecting real-world data challenges. This training dataset forms the basis for machine learning model development, with a separate test dataset reserved for performance evaluation. Discrepancies in the ratio of PCR-positive to PCR-negative instances add complexity, requiring advanced predictive modelling techniques. The 'pCR (outcome)' feature serves as the classification model's target variable among 119 features, while 'RelapseFreeSurvival (outcome)' is designated for the regression model. Recognizing machine learning's potential in this context is crucial for enhancing breast cancer treatment precision and efficacy through data-driven insights.

3. DATA CLEANING AND PREPROCESSING

Several procedures were performed during the data cleaning and preparation phase to ensure the dataset's quality and applicability for machine learning research. Further examinations against redundant information, inconsistencies, and data types showed no duplications, inconsistencies, or numeric data. These preprocessing methods enhanced the dataset's quality and prepared it for the machine learning development of models.

3.1. Drop Columns

To begin with, the 'ID' field was dropped from the dataset because it was deemed of little significance for predicting outcomes. Since each data point has a unique "ID", an "ID" column does not add to these patterns. Additionally, due to its categorical nature, it may raise the dimensionality of the feature space without contributing any additional benefit when one-hot encoded.

3.2. Missing Values

A mean replacement technique was used to address missing values designated as '999,' and the resulting numbers were rounded to integers in order to maintain the originality of the dataset. By replacing missing numbers with the mean, you are effectively replacing them with a value that accurately reflects the data's central tendency. This retains the general distribution and introduces no major bias.

3.3. Outliers

Outliers were managed through z-score normalization, excluding 'pCR (outcome)' and 'RelapseFreeSurvival (outcome)' as they served as primary target variables. Utilizing z-scores within the range of -3 to 3 for outlier

identification offers a standardized, straightforward, and quantifiable method. This approach maintains uniformity, simplicity in implementation, and a measurable indication of data points' deviation from the mean, robust to skewness.

3.4. One- Hot Encoding

To accommodate multi-class records within a specific column, relevant columns were encoded using One-Hot Encoding technique. By transforming categorical data into a binary format, One-Hot Encoding retains the distinctiveness of categories, minimizing ordinal misinterpretation, and strengthening compatibility with different machine learning techniques. List of categorical variables- ‘ER', 'PgR', 'HER2', 'TrippleNegative', 'ChemoGrade', 'Proliferation', 'HistologyType', 'LNStatus', 'TumourStage'.

3.5. Normalization

Normalization is pivotal in machine learning preprocessing to ensure consistent feature scales, prevent variable dominance, aid algorithm convergence, enhance model flexibility, and improve the performance of algorithms sensitive to varying feature scales. The dataset underwent normalization using StandardScaler, with the exclusion of target variables. The application of Standard Scaler is chosen for normalization due to its promotion of scale consistency, compatibility with various algorithms, expedited convergence, optimized interpretability, robustness to outliers, facilitation of regularization, adherence to statistical assumptions, reduced complexity in hyperparameter tuning, and enhanced model performance.

3.6. Handling Class Imbalances

Finally, to eliminate any potential class imbalances in classification models, the oversampling issue was addressed using Random Oversampling. As oversampling produces more representative data for minority groups, it reduces bias and improves pattern capture without additionally wasting information. In case of classification, the pre-processed dataset contains 84 for target class 1 and 316 of target class 0 . Since the dataset is imbalanced, oversampling method is used to balance.

3.7. X and y Split

The normalised dataset was subsequently split into target and predictor variables, with 'pCR (outcome)' and 'RelapseFreeSurvival (outcome)' serving as the goal parameters (X) for classification and regression models, respectively, with the remaining variables serving as predicting factors (y).

4. FEATURE ENGINEERING

In the feature engineering phase, a systematic approach was taken to enhance the predictive capabilities of the dataset. This was conducted using the embedded method, Dimension reduction and finally train-test split to facilitate model evaluation. These feature engineering strategies collectively contribute to refining the dataset and optimizing its suitability for subsequent machine learning model development and evaluation.

4.1. Feature Selection

Feature selection was conducted using the embedded method known as The Least Absolute Shrinkage and Selection Operator (LASSO). This technique, particularly effective with numerous features, automatically selects features by minimizing the error sum of squares through coefficient penalization. In this study, LASSO Regression utilized a hyperparameter alpha set to 0.01 for balanced feature selection. Table [4.1.1] demonstrates a comparison of Lasso, Ridge, and Elastic Net. Lasso excelled in effectively minimizing features, while Ridge retained features and Elastic Net yielded zero columns. Consequently, Lasso was chosen for feature selection. Additionally, Lasso (L1 regularization) is often preferred over Ridge and Elastic Net as it incorporates an inherent feature selection process, promoting sparsity by driving certain coefficients to zero, resulting in simpler and more interpretable models.

Feature Selection Method	Parameters	Selected features count
LASSO	alpha=0.01	41
RIDGE	alpha=1.0	130
ELASTIC NET	alpha=1.0, l1_ratio=0.5	0

[4.1.1] Comparison between Feature Selection Methods

4.2. Train and Test Split

The processed dataset underwent a train-test split to facilitate model evaluation. The hyperparameter test size was set to 0.2, indicating a 20% allocation for testing data, while the remaining 80% constitutes the training set. Also, random state was fixed at 42 to ensure result reproducibility.

5. MODEL SELECTION

In model selection, three models were chosen for each classification and regression task. This deliberate choice includes a mix of models suitable for both linear and non-linear data, ranging from simpler to more advanced models. This diverse selection aims to facilitate a thorough comparison, shedding light on the performance nuances across different data scenarios for a comprehensive understanding.

5.1. Classification

In the model selection process, three classifiers—AdaBoostClassifier, DecisionTreeClassifier, and LogisticRegression—were assessed based on their balanced accuracy and corresponding parameters. AdaBoost combines shallow decision trees sequentially to improve predictive accuracy, especially useful when individual models perform sub optimally. Decision Trees, known for their hierarchical structure, make binary decisions at each node and capture intricate non-linear relationships. Logistic Regression models the probability of an instance belonging to a class by fitting a logistic function to input features. Together, these methodologies play a crucial role in addressing diverse classification challenges in academic research.

5.2. Regression

In the regression analysis, Support Vector Regression (SVR), Gradient Boosting, and LASSO Regression were applied. Similar to Ada Boosting, Gradient Boosting employs weak decision trees but prioritizes error correction by modeling residuals, distinguishing it from Ada Boost's instance weighting based on misclassification. Support Vector Regression (SVR) aims to identify an optimal hyperplane representing the data, minimizing errors through margin optimization. Notably, our study observed superior performance with the linear configuration of SVR. LASSO Regression, introduces regularization by penalizing the absolute values of the regression coefficients. Together, these regression techniques, chosen for their diverse approaches, contribute to a comprehensive exploration of predictive modelling in our research, incorporating nuances from both error correction and regularization.

6. HYPERPARAMETER TUNING

GridSearchCV is employed for hyperparameter tuning in both the classification of Pathological Complete Response (PCR) and the regression of Recurrence-Free Survival (RFS). It is selected for its ability to thoroughly assess all combinations in a predefined grid, ensuring comprehensive evaluation through cross-validation.

For classification, Grid search is conducted with K-fold validation, where K is set to 5. The scoring metric employed is balanced_accuracy.

$$grid\_search = GridSearchCV(classifier, param\_grid, cv = 5, scoring = 'balanced\_accuracy', n\_jobs = -1)$$

Model	Balanced Accuracy	Best Parameters
AdaBoostClassifier	86.9%	{'algorithm': 'SAMME.R','base_estimator': DecisionTreeClassifier('learning_rate': 1.0, 'n_estimators': 200}
DecisionTreeClassifier	89.7%	{'criterion': 'gini', 'max_depth': 45, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 2, 'splitter': 'best'}
LogisticRegression	70.5%	{'C': 1, 'fit_intercept': False, 'max_iter': 50, 'penalty': 'l2', 'solver': 'lbfgs'}

6.1.1. Results from GridSearchCV for classification.

Table [6.1.1] demonstrates the robust performance of the decision tree model. The efficacy of Ada Boost is found to be marginally diminished. This is attributable to Ada Boost's reliance on an ensemble of decision trees with varying individual performances. Notably, the decision tree exhibits superior performance compared to logistic regression, primarily attributed to its inherent capacity to capture nuanced non-linear relationships within our dataset.

For regression, Grid search is executed with K-fold validation, where K is set to 5. The scoring metric utilized is the balanced 'neg_mean_absolute_error.'

$$grid\_search = GridSearchCV(regressor, param\_grid, cv = 5, scoring = 'neg\_mean\_absolute\_error', n\_jobs = -1, verbose = 2)$$

Model	MAE (Mean Absolute Error)	Parameters
SVR(Support vector regression)	22.96	{'C': 0.1, 'epsilon': 0.1, 'gamma': 'auto', 'kernel': 'linear'}
GradientBoostingRegressor	20.78	{'learning_rate': 0.01, 'loss': absolute_error, 'max_depth': 4, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 400, 'subsample': 0.8}

LASSO	23.41	{'alpha': 1, 'fit_intercept': True, 'max_iter': 100}

6.2.1. Results from GridSearchCV for regression.

Based on Table [6.2.1], Gradient Boosting Regression consistently surpasses SVR and Lasso Regression in performance, attributed to its adept handling of non-linear relationships, adaptability to complex patterns, and resilience against outliers. While SVR and Lasso Regression exhibit strengths in certain contexts, the ensemble approach of gradient boosting renders it versatile and potent across a diverse array of regression problems. The sequential error correction during training further enhances its predictive prowess.

7. K-FOLD VALIDATION – MODEL VALIDATION

K-fold validation is employed for robust model validation, ensuring a thorough assessment of performance in Pathological Complete Response (PCR) classification and Recurrence-Free Survival (RFS) regression. This approach optimally uses available data, mitigating overfitting and enhancing generalization to new data. The use of K-fold validation aids in visualizing training and validation metrics, facilitating the identification of overfitting and underfitting.

For classification tasks, K-fold validation with Cv=5 is implemented, utilizing the balanced_accuracy metric for scoring. In regression scenarios, 'cv=5' is employed, with mean absolute error (MAE) serving as the scoring metric.

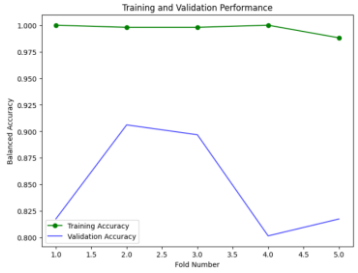

Fig [1] Model Performance for Classification - Decision Tree

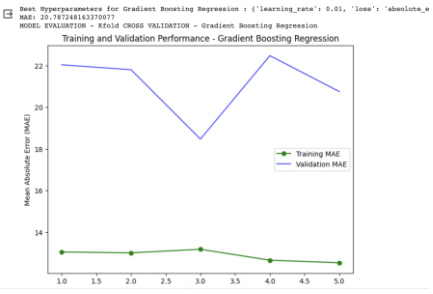

Fig [2] Model Performance for Regression - GradientBoosting

Fig [1], Fig [2] aid in assessing the model's generalization to unseen data across folds. The plots offers insights into the model's performance consistency across diverse subsets of the data.

8. PREDICTING USING THE TEST SET

The trained classifier's predict method generates predictions for the test set's feature matrix (X_test). The resulting array (y_pred) with predicted labels is then compared to the true labels (y_test) for performance evaluation.

The balanced_accuracy_score is calculated to be 95.24% using the following formula:

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{True\ Positive\ Rate(Sensitivity)}{Positive\ Class} + \frac{True\ Negative\ Rate(Specificity)}{Negative\ Class}\right)$$

Balanced accuracy has been chosen due to its robustness in predicting with test sets, particularly in imbalanced datasets. It ensures a fair evaluation of the classifier's performance across all classes, thereby enhancing reliability in real-world scenarios characterized by uneven class distributions.

Classification	
Model	Balanced Accuracy score
AdaBoost	94.4%
DecisionTree	98.4%
Logistic Regression	75.9%

Regression	
Model	MAE
GradientBoost	13.16
SVR	21.25
LASSO	20.72

8.1.1. Test Prediction Results for Classification and Regression.

According to Table [8.1.1], the decision tree yielded the highest performance with a balanced accuracy of 98.4%, while gradient boosting regression exhibited the lowest Mean Absolute Error (MAE) at 13.16. Consequently, the decision has been made to employ the decision tree for predicting PCR outcomes and gradient boosting regression for forecasting RFS results.

9. CONCLUSION

This research focuses on utilizing machine learning to detect breast cancer using a dataset of 400 patients with clinical and MRI features. The approach involved robust pre-processing, addressing imbalances, and feature selection via LASSO regression. Decision Tree excelled in classification, while Gradient Boosting Regression outperformed in regression tasks. Hyperparameter tuning and cross-validation optimized our models, emphasizing the efficacy of our approach for accurate breast cancer prediction.

10. REFERENCES

1. "Radiomics based likelihood functions for cancer diagnosis," nature.com. https://www.nature.com/articles/s41598-019-45053-x#Sec4 (accessed Dec. 3, 2023).

11. CONTRIBUTION PERCENTAGE

Task and Weighting	Data preprocessing (10%)	Feature Selection (25%)	ML method development (25%)	Method Evaluation (10%)	Report Writing (30%)
Harnitha Suresh	20%	20%	20%	20%	20%
Shruti Sundaram	20%	20%	20%	20%	20%
Kunal Choudhary	20%	20%	20%	20%	20%
Omkar Mainkar	20%	20%	20%	20%	20%
Tejas Chavan	20%	20%	20%	20%	20%